APPLIED ECONOMETRICS

Potluri Rao Roger LeRoy Miller

University of Washington

Wadsworth Publishing Company, Inc., Belmont, California

1971

Chapter 2

Uses of Summary Statistics in Linear Regression

After having obtained estimates of the parameters in a linear regression equation, the researcher usually computes summary statistics to assess the usefulness of the estimates. The meaning and use of these statistics is especially valuable in applied econometrics, where their role is not always to provide cut-and-dried answers. Instead they are the basic tools of the applied econometrician in working his way through an empirical problem. The typical "textbook" econometric problem in which the model is clearly specified is seldom encountered in empirical research. Applied econometricians use a great deal of judgment at various stages of research by utilizing summary statistics to "feel the data."

It is difficult to set up definite rules concerning the uses and abuses of these summary statistics as econometric tools, since their proper selection requires skill and intuition on the part of the researcher. He should be aware of when to use a tool and also of when not to use it. It should be kept in mind throughout this book that applied econometrics is an art.

2.1 Multiple Correlation Coefficient

In linear regression estimation, the residuals indicate the extent of the movement in the dependent variable that is not explained by the independent variables. If the residuals are small relative to the total movement in the dependent variable, then it follows that a major part of the movement has been accounted for. Accordingly, the summary statistic known as the multiple correlation coefficient is defined to measure the extent of movement in the dependent variable that is explained by the independent variables. Conventionally, instead of the multiple correlation coefficient, its square (R^2) is reported with all the regression equations.

The square of the multiple correlation coefficient is defined as

$$R^2 = \frac{variation\ explained\ by\ the\ regression\ equation}{total\ variation\ of\ the\ dependent\ variable} . \qquad (2.1)$$

Since the residuals represent the movement in the dependent variable that is unexplained by the independent variables, the R^2 may also be expressed as

$$R^2 = \frac{\Sigma(Y-\bar{Y})^2-\Sigma e^2}{\Sigma(Y-\bar{Y})^2} \qquad (2.2)$$

where $(\Sigma(Y-\bar{Y})^2)$ and (Σe^2) are variations of the dependent variable and of the residual respectively.

The summary statistic so defined measures the proportion of variation in the dependent variable that is explained by the independent variables. This particular definition has a valid interpretation as a summary statistic only when the regression equation contains the constant term either explicitly or implicitly.

In ordinary least squares estimation, when the researcher includes an additional variable in the regression equation the sum of squares of the residuals (Σe^2) necessarily decreases. This is a mathematical property and does not depend on the "relevance" of the additional variable in the causal relation. Therefore, whenever a variable is added to the regression the R^2 necessarily increases.

The researcher should note that even though R^2 is used as a measure of the proportion of variation in the dependent variable that is explained by the regression equation, it should not always be interpreted as a determinant of "goodness of fit" of the causal relation. We will treat this point in more detail later on.

As an illustration, consider the consumption function estimated for the United States for the time period 1956 to 1960.

$$C_t = -0.34 + 0.76\, Y_t + 0.30\, C_{t-1} \qquad (2.3)$$

The sum of squares of residuals (Σe^2) corresponding to this equation is 0.0015, and the variation in the dependent variable $(\Sigma(C-\bar{C})^2)$ is 0.3010. Using definition (2.2), the R^2 may be computed as

$$R^2 = \frac{0.3010-0.0015}{0.3010} = 0.99\,. \qquad (2.4)$$

When presenting his results, the researcher customarily gives the statistic R^2 as a part of the estimated regression equation, usually in the following manner:

$$C_t = -0.34 + 0.76\,Y_t + 0.30\,C_{t-1} \qquad R^2 = 0.99\,. \qquad (2.5)$$

Regression equation (2.5) explains 99 percent of the variation in the dependent variable *C*.

Suppose that instead of estimating the consumption function the researcher is trying to explain aggregate savings, and suppose that he has postulated the causal relation

$$S_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 C_{t-1} + \varepsilon_t\,, \qquad (2.6)$$

where *S*, is aggregate savings in a given quarter defined as

$$S_t = Y_t - C_t\,. \qquad (2.7)$$

For the data 1956 to 1960 the researcher would obtain the following estimated equation:

$$S_t = 0.34 + 0.24\,Y_t - 0.30\,C_{t-1} \qquad R^2 = 0.64 \qquad (2.8)$$

Whereas the consumption function explains 99 percent of the variation in the dependent variable, the savings function explains only 64 percent. The researcher, concluding that the consumption function is a “better causal relation” than the savings function, may be tempted to retain the consumption function for policy purposes and never to mention the savings function again. He should cultivate resistance to such temptations, for reasons that will become clear.

Let us turn now to an interpretation of the consumption and savings functions. If while holding the previous quarter’s consumption constant the income of the economy (*Y*) is increased by one dollar, then according to the consumption function the consumption in the current quarter increases by 0.76 dollars, and according to the savings function savings increase by 0.24 dollars—which is, in fact, that part of the additional dollar that was not spent (1 - 0.76) according to the consumption function. Both the consumption function and the savings function are providing identical information on the consumption behavior of the United States. Similarly, holding current income constant, if we increase the previous quarter’s consumption by one dollar we get identical answers from both equation (2.5) and equation (2.8).

Even though the two regression equations corresponding to the same data are providing identical information on the economy, one has a larger value for the summary statistic R^2 than the other. To use R^2 as a measure of the appropriateness of the regression equation for explaining the movements in the dependent variable would, in this situation, be a

misuse of the statistic.

A high R^2 may imply the appropriateness of a regression equation for explaining the movements of a dependent variable, but a low R^2 does not necessarily imply that the regression equation is inappropriate. The source of this possible misuse of R^2 lies in its definition.

In estimating the two equations (2.5) and (2.8) the researcher is imposing certain conditions on the residuals of each. Close examination reveals that these conditions are the same for both estimations. Since the same set of residuals satisfies the conditions imposed by the two equations, the implicit regression estimates and the variation in the residuals must be the same. The implicit correspondence between the regression coefficients may readily be seen by inserting the identity (2.7) into the savings function (2.8) to produce

$$Y_t - C_t = 0.34 + 0.24\,Y_t - 0.30\,C_{t-1} \qquad (2.9)$$

which is nothing but the consumption equation (2.5) expressed in a different form. This can be seen by taking *Y*, to the right-hand side of (2.9) and then multiplying through by -1. The result is (2.5).

Even though the residual sum of squares for the two regression equations is the same, the variation in the dependent variables is different because in one case we have *C* and in the other case *S* as the dependent variable: $\Sigma(C_t - \bar{C})^2$ is not equal to $\Sigma(S_t - \bar{S})^2$. By using definition (2.2) we obtain different values for R^2 for the two equations even though the implicit regression coefficients and the variation in the residuals are identical.

In general, if instead of using the dependent variable one uses any linear combination of dependent and independent variables, he is bound to get the same residual sum of squares and implicit parameter coefficients, but different R^2's. Numerous examples can be found in current empirical studies. The most common case occurs in demand functions, in which the quantity demanded is specified as a function of certain independent variables. Given a change in some independent variable, a new value of the dependent variable is obtained, but this “desired” quantity cannot be attained instantaneously.

Therefore, the partial-adjustment mechanism is posited. A typical resulting equation might then be

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 Y_{t-1} + \varepsilon_t \ . \qquad (2.10)$$

Suppose that instead of the *level of quantity demanded*, Y_t, the researcher wants *changes in demand* to be predicted by his regression equation. He then subtracts Y_{t-1} from both sides of the equation, obtaining

$$(Y_t - Y_{t-1}) = \beta_0 + \beta_1 X_{1t} + \beta_3 Y_{t-1} + \varepsilon_t \ . \qquad (2.11)$$

The R^2 from this equation will usually be smaller than the R^2 from the equation (2.10), but a relatively low R^2 does not necessarily mean a poor fit. Note that we are explaining the variance of *changes* in Y_t in one equation and the variance of Y_t in the other.

This result occurs because the dependent variable is not the same in both equations. Another situation arises when the dependent variables are different functional forms of the same variable. In these cases, also, the R^2 cannot be used for comparison of the two equations. Consider, for example, the following:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_{1t} \qquad (2.12)$$

$$\log Y_t = \gamma_0 + \gamma_1 X_{1t} + \gamma_2 X_{2t} + \varepsilon_{2t} \ . \qquad (2.13)$$

The specification of the model, the error terms, and the computation of R^2 for these two equations are entirely different and provide no common ground for comparison of the relative performance of these equations on the basis of computations of R^2.

It is clear from the above discussion that R^2 can be legitimately used for comparison of the relative performance of two competing regression equations only when the dependent variables are the same. And since R^2 always increases when independent variables are added, we need the further restriction that the number of *X*'s be the same in each equation being compared.

It may seem now that R^2 has little, if any, value; but this is not the case. For example, we may use it to determine which of several competing definitions of an independent variable is most appropriate empirically.

Consider a simple Cobb-Douglas production function:

$$\log Q = \beta_0 + \beta_1 \log K + \beta_2 \log L + \varepsilon \ . \qquad (2.14)$$

The variable *K*, capital, in this example is assumed to be well defined, but suppose the researcher is faced with several competing definitions of the variable *L*, labor. Suppose also that on theoretical grounds it is not obvious which of the several definitions is most

appropriate. One way to proceed is to estimate a separate production function for each of the definitions of labor; the definition that produces the highest R^2 when used in the equation may be considered empirically preferable. The argument for this procedure is that the precise empirical definition of variables should be selected so as to put the theory in question in its best light.

This procedure should not be misused. It applies only to the choice among a well-selected and theoretically acceptable set of alternative definitions of a given variable.

It may happen that a nonsensical definition of the variable will give the highest R^2; this, of course, does not mean that it is the appropriate one to use. Basing the choice of appropriate definition of an independent variable on a maximum R^2 is justified only when the model has been fully specified and all the other variables of the model are well defined. This procedure is a guide in empirical research, and not a theoretical rule.

Consider the case of a researcher who is working on the same regression equation for several sets of data—for example, the production function for each state separately. It may be that no definition of a particular variable, say labor, will give the highest R^2 in every case. Suppose the theoretically valid definition is efficiency units of labor, and suppose that data are available on variables such as the number of workers, wages paid, education of workers, industrial concentration, etc. The efficiency of workers may depend in one state on the level of education, in another state, on the concentration of minority groups, etc. In these cases selection according to the highest R^2 will yield different empirically appropriate definitions of labor. When use of the highest R^2 suggests the use of different definitions for different states then, of course, the researcher should investigate the reasons underlying such behavior of the data. When a theoretical justification for the anomaly is not forthcoming and when he has strong reasons to believe that the efficiency units of labor in all the states should have the same definition, then he may select the definition that gives the highest R^2 most of the time, using his own judgment in weighing empirical and theoretical considerations underlying the problem.

Whenever he uses R^2 as a criterion the researcher should guard against inadvertently selecting a possible candidate which is actually some disguised form of the dependent variable. For example, suppose that the researcher is estimating the relation between output and man-hours. The published monthly and quarterly output data may have been interpolated by the data-collecting agency from information on man-hours. Thus the researcher should expect to obtain a high R^2 when he regresses output on man-hours or man-hours on output. A high R^2 in this case gives little, if any, information.

When the R^2 values from different definitions of a variable do not differ substantially from each other, then the use of the highest R^2 as a guide has little significance, and one definition is as good as the other.

Another example of the meaningful use of highest R^2 concerns the choice of the appropriate lag for an independent variable. Suppose that we wish to explain the level of interest rates, i_t , and suppose we have theoretically considered the lagged money supply to be a relevant independent variable along with many others: That is, a change in the money supply is considered to have a once-for-all effect on the level of the "interest rate" after τ time periods. (This is not to be confused with the distributed lag effect to be discussed in Chapter 7.) Then our regression equation would be specified as

$$i_t = \beta_0 + \beta_1 M_{t-\tau} + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t \ , \qquad (2.15)$$

where τ is the lag period in the money supply (*M*).

The empirically appropriate lag in the supply of money in explaining the interest rate may be obtained by fitting equation (2.15) for various values of τ and considering as empirically appropriate that value which gives the highest R^2. This procedure is, of course, valid only when the rest of the specification is correct.

In empirical research one comes across many such occasions where R^2 can be used as a guide rather than as a summary statistic.

2.2 Residual Variance

Consider the case in which the dependent variable is the same in all regressions, but the list of independent variables is different. For example, examine the following two regression equations:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_{1t} \ , \qquad (2.16)$$

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_{2t} \ . \qquad (2.17)$$

Even though the dependent variables are the same, the R^2 for the two equations are not comparable because the number of independent variables is not the same. As mentioned before, the second equation (2.17) necessarily gives a higher R^2. The sum of squares of the residuals will be smaller, but estimation of the second equation imposes an extra condition on the residuals. Is the reduction in the residual sum of squares worth the *price* of the extra constraint? To answer this, a summary statistic—residual variance—

is computed:

$$V(e) = \Sigma e^2/\upsilon \ , \qquad (2.18)$$

where υ is the degrees of freedom (that is, the total number of observations less the number of constraints imposed on the residuals in estimating the parameters). Even though the residual sum of squares (Σe^2) necessarily decreases with the addition of a variable, the residual variance, V(e), need not, because the denominator in (2.18) is also changing. The residual variance takes into consideration information about the degrees of freedom, whereas the R^2 does not. Note that V(e) as defined in equation (2.18) is meaningful only when the regression equation has a constant term.

A summary statistic analogous to R^2 is defined on the basis of the residual variance by

$$\bar{R}^2 = 1 - V(e)/V(Y) \ , \qquad (2.19)$$

where $V(Y)$ is the variance of Y defined as $V(Y) = \Sigma(Y-\bar{Y})^2/(T-1)$. So defined, the statistic $\bar{R}^2$ can decrease when a new variable is added to a regression equation even though R^2 necessarily increases. Since $V(Y)$ does not depend on the independent variables, there is one-to-one correspondence between the $\bar{R}^2$ and the variance of the residual V(e).

One should not jump to the conclusion that the equation which yields least residual variance (the largest $\bar{R}^2$) is necessarily always desirable. In a regression equation, the decision on including or excluding a variable is based on theoretical considerations and the use to which the regression is put, rather than on mere maximization of the summary statistics R^2 and $\bar{R}^2$.

However, as an example of an occasion when the equation with the least residual variance is desirable, consider the case in which a researcher is interested in predicting the values of a dependent variable to a yet-unknown period. The predicted value need not be the same as the value of the dependent variable actually observed for that period, because of the error of prediction. Given two predictors, one would choose that one which has the smallest variance of the error of prediction. A regression equation with smaller residual variance also has smaller variance of the error of prediction.

When adding an independent variable increases the $\bar{R}^2$, the prediction power can be increased by including that variable, because the variance of the error of prediction is thereby decreased. When adding an independent variable decreases the $\bar{R}^2$, then the researcher, of course, is losing reliability in prediction by including that variable in the regression.

When the objective is testing of a null hypothesis based on the regression estimates and not prediction of a future value of the dependent variable, then the researcher is interested in unbiased estimates of his parameters. Unbiased estimates may be obtained only by including all the theoretically specified variables in a regression equation irrespective of what they do to the summary statistic $\bar{R}^2$. Discarding a theoretically relevant variable may increase $\bar{R}^2$ but it may result in biased estimates of the parameters.

The problem of prediction is an integral part of econometric research. When a researcher is interested primarily in predicting values of a specific variable rather than in testing a theory, he will choose his variables in such a way as to obtain the regression equation with the least residual variance.

2.3 Standard Errors

Having obtained the regression coefficients by using the ordinary least squares procedure, the researcher is interested in assessing the "precision" of the estimation procedure. To this end, he computes standard errors of the regression coefficients.

These computed standard errors do not necessarily reflect the true precision of the estimates, which is measured by the theoretical variance of the distribution of an estimate, not by the standard errors. The theoretical variance, however, is an unknown. To bring out the distinction let us first consider the theoretical variance.

When several estimation procedures exist for the same parameters from the same data, the researcher would like to use the one offering maximum precision. Even when no choice in estimation procedures is available, he may still be interested in discovering the precision of his estimates for the purpose of testing a null hypothesis.

Consider a situation in which the truth is

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t , \qquad (2.20)$$

where lower-case letters indicate deviations from the mean; hence the constant is implicit. Let us suppose that the values of the parameters (β's) and the independent variables (x's) are known. Then the value of the dependent variable depends on the values of the error terms in equation (2.20). Since in this example all the terms on the right-hand side of the equation are known except for the error terms, the values of y vary with the error terms alone. For different sets of y's (which depend on the $\epsilon's$), and of

given *x*'s, we obtain different estimates of the parameters by using the ordinary least squares estimation procedure. The question then is: how sensitive are the estimates going to be to the particular values of the error terms?

We are assuming that the error terms are distributed independently of each other and that they are uncorrelated with the independent variables (*x*'s). Many sets of error terms may satisfy these conditions. Suppose the errors are a particular set, say ε^*. Corresponding to these error terms is a set of *y* values given by equation (2.20). When these values of *y* are regressed on the independent variables we obtain one set of estimates for the parameters, say $\hat{\beta}^*$. These estimates need not be equal to the parameters, because the ordinary least squares estimates are chosen to minimize the residual sum of squares, and the true parameter values need not correspond to the minimum residual sum of squares in any given sample. If, instead of ε^*, the errors were a different set, say ε^{**}, then we would have obtained a different set of estimates for the parameters, say $\hat{\beta}^{**}$. Since we do not know which set of errors corresponds to the given data on the *y*'s, we cannot say anything about the extent of deviation of these estimates from the true parameters. We can, however, study how much dispersion the estimates would exhibit if the errors were in fact drawn at random from a population with known variance. In such a situation any combination of error terms has an equal chance of corresponding to the given data on the *y*'s. When the errors are drawn at random from a population with zero mean and constant variance σ_ε^2, then the regression coefficients have a distribution which can be established analytically, as will be shown in Chapter 3.

In defining the precision of the estimates in this way, we are using information on the distribution of error terms when the values of independent variables and parameters are known. Since the precision is defined conditionally upon these values, we expect precision to depend on these values as well. In a linear regression equation the precision depends only on the values of the *x*'s, and not on the values of the parameters.

The conventional way of measuring precision when the statistical distribution of an estimate is known is by its variance. The smaller the variance of an estimate the greater its precision—that is, the less sensitive the estimates will be to different sets of error terms that could have occurred in the *y*'s.

The theoretical variance of the distribution of $\hat{\beta}_1$ in equation. (2.20) may be derived as

$$V(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\Sigma x_1^2(1-r_{x1x2}^2)} \qquad (2.21)$$

where σ_ε^2 is the variance of the population that generated the error terms and r_{x1x2}, is the

correlation between *x*1 and *x*2. It can be seen from equation (2.21) that the larger the variation of the independent variable *x*1 relative to the variance of the error term, the smaller the variance of the estimate $\hat{\beta}_1$. That is, the precision of the regression coefficient corresponding to an independent variable increases with the variation of that variable. Precision also depends on the co-movements of the independent variables. The smaller the correlation between the independent variables, the higher the precision of the regression estimates.

These are theoretical results, and they involve σ_ε^2, which is generally unknown. The researcher can, however, estimate the variance of the error term from the residuals of the regression equation. And, since the variance of the estimate of $\hat{\beta}_1$ is unknown, we may estimate it by replacing σ_ε^2 by its sample estimate. Thus, an estimate of the variance of $\hat{\beta}_1$ is

$$\hat{V}(\hat{\beta}_1) = \frac{V(e)}{\Sigma x_1^2(1-r_{x1x2}^2)} , \qquad (2.22)$$

where $\hat{V}(\hat{\beta}_1)$ stands for an estimate of V($\hat{\beta}_1$), and V(e) is the sample variance of the residual (e):

$$V(e) = \Sigma e^2/\upsilon \quad \text{where } \upsilon = \text{degrees of freedom.} \qquad (2.23)$$

The variance is in square units of the regression coefficient. To convert them to comparable units, the standard deviation of the regression coefficient is defined as the square root of the variance of the regression coefficient. The true standard deviation is a theoretical quantity. When computation of the standard deviation is based on the *estimate of variance* rather than on the *variance itself*, it is called the *standard error* to distinguish it from its theoretical value.

Whenever the researcher reports the standard error of his estimate he is explicitly stating that the result is an estimate of the standard deviation of the coefficient and is not the standard deviation itself. Despite the distinction in labels, researchers often overlook this point and use standard errors as if they were standard deviations; such misuse of the results should be discouraged.

When the estimate of variance of the regression coefficient is biased, the estimated variance does not reflect the precision of the regression coefficients. In such cases the theoretical standard deviation reflects the precision, but the standard error does not.

When the researcher has a choice between two alternative estimation procedures, he may wish to choose the one having the greater precision. Since the precision of estimates is reflected only in the standard deviations and not in the standard errors, he

should not conclude that the estimation procedure which yields smaller standard errors is necessarily preferable. A thorough investigation into the theoretical properties and standard deviations must precede such decisions.

Computational precision as evidenced by small standard errors of regression coefficients does not necessarily indicate that the most theoretically precise estimation procedure has been used.

To illustrate the above points, consider the consumption function example:

$$C_t = \hat{\beta}_0 + \hat{\beta}_1 Y_t + \hat{\beta}_2 C_{t-1} + e_t \ , \qquad (2.24)$$

where the $\hat{\beta}$'s are the least squares estimates. The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ may be computed from the following summary information:

$$V(e) = \frac{\Sigma e^2}{\upsilon} = \frac{0.001512}{19-3} = 0.0000945 \ , \qquad (2.25)$$

$$\Sigma y_t^2 = \Sigma(Y_t - \bar{Y})^2 = \Sigma Y_t^2 - \Sigma Y_t . \Sigma Y_t / T = 0.2604 \ , \qquad (2.26)$$

$$\Sigma c_{t-1}^2 = \Sigma(C_{t-1} - \bar{C})^2 = \Sigma C_{t-1}^2 - \Sigma C_{t-1} \Sigma C_{t-1} / T = 0.2845 \ , \qquad (2.27)$$

$$r_{y_t c_{t-1}} = \frac{\Sigma y_t c_{t-1}}{\sqrt{\Sigma y_t^2 . \Sigma c_{t-1}^2}} = 0.9728 \ . \qquad (2.28)$$

To obtain a measure of precision, we should use the formulae for theoretical variance:

$$V(\hat{\beta}_1) = \sigma_\varepsilon^2 / \Sigma y_t^2 (1 - r_{y_t c_{t-1}}^2) \ , \qquad (2.29)$$

$$V(\hat{\beta}_2) = \sigma_\varepsilon^2 / \Sigma c_{t-1}^2 (1 - r_{y_t c_{t-1}}^2) \ . \qquad (2.30)$$

Since σ_ε^2 is unknown, we shall estimate the variances by replacing σ_ε^2 by V(e):

$$\hat{V}(\hat{\beta}_1) = 0.0000945/(0.2604)(0.0272) = 0.00676, \qquad (2.31)$$

$$\hat{V}(\hat{\beta}_2) = 0.0000945/(0.2845)(0.0272) = 0.00619 \ . \qquad (2.32)$$

The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ are therefore

$$St.error(\hat{\beta}_1) = \sqrt{\hat{V}(\hat{\beta}_1)} = 0.082 \ , \qquad (2.33)$$

$$St.error(\hat{\beta}_2) = \sqrt{\hat{V}(\hat{\beta}_2)} = 0.079 . \qquad (2.34)$$

It is conventional to report the standard errors of the coefficient estimates as a part of the regression equation by enclosing them in parentheses below the respective estimates. In its final form of presentation the consumption function would be

$$C_t = \underset{}{-0.34} + \underset{(0.082)}{0.76}\, Y_t + \underset{(0.079)}{0.30}\, C_{t-1} \qquad R^2 = 0.99 . \qquad (2.35)$$

In the case of more than two variables, computation of standard errors by using explicit formulae as above would be computationally burdensome. Standard computer programs provide the standard errors of the estimates, together with the estimates, at an insignificant marginal cost.

2.4 Bias in the Estimates

For any given set of independent variables the regression estimates of a parameter depend on the error terms actually present in the data. When the errors are not known but are assumed to have been drawn randomly from a population with known distributional properties, then the regression estimates have a statistical distribution. We studied the interpretation of the variance of this distribution in the preceding section. Now we concentrate on the mean of this distribution.

The statistical distribution of a regression estimate may or may not center around its corresponding parameter. Consider, for example, a case in which there are three different alternative ways of estimating a parameter, β, and in which the distributions of these estimates ($\hat{\beta}$, $\widetilde{\beta}$, and β^*), are as presented in Figure 2.1.

Such estimates are called *unbiased*. The estimates $\hat{\beta}$ and $\hat{\beta}^*$ are not centered around their parameter values and are called *biased*. The bias of an estimate is measured as the difference between the mean value of the distribution of an estimate and its true parameter value. When the bias is positive—that is, when the mean value of the distribution is larger than its parameter, as in the case of $\hat{\beta}^*$—then the estimator is said to be upward biased. Conversely, when the bias is negative, as in the case of $\hat{\beta}$, the estimator is negatively biased.

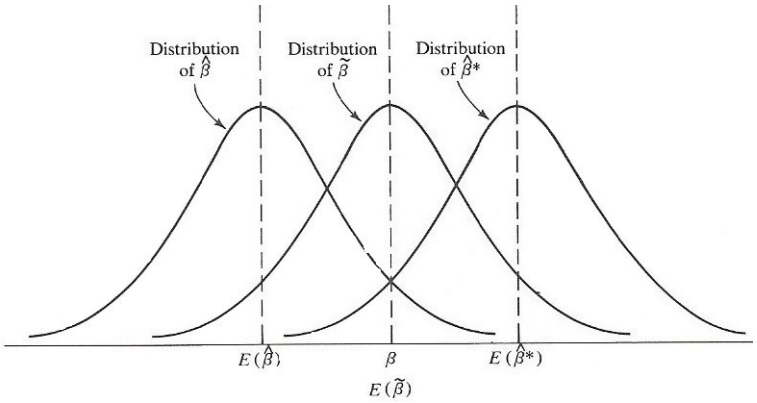

Figure 2.1. Distribution of $\hat{\beta}$, $\tilde{\beta}$, and $\hat{\beta}^*$

The distribution of $\hat{\beta}$ is centered around its parameter value β.

The use to which an estimate is to be put determines which one is chosen as "appropriate." When the researcher is interested in testing a null hypothesis, for example, he may prefer an unbiased estimate because many test procedures are based only on unbiased estimators of a parameter. When his objective is not the elimination of bias but maximum precision, then, of course, he will look for the minimum variance estimator even though it may be biased.

When the researcher is faced with a situation in which he has to choose between an unbiased estimate with large variance and a biased estimate with small variance, he cannot use either the bias or the variance as the sole criterion but must give some weight to each aspect. One criterion that gives equal importance to these two measures is the mean square error. Since the variance is in the square units of bias, the giving of equal importance to bias and to variance implies that equal importance is attached to variance and to the square of bias. Thus the mean square error is defined as the sum of variance and of the square of bias. If one of these two components gets smaller at the expense of the other, then the net "benefit" is reflected by the mean square error. The mean square error (MSE) of an estimate $\hat{\beta}$ may be written as

$$MSE(\hat{\beta}) = V(\hat{\beta}) + [Bias(\hat{\beta})]^2 . \qquad (2.36)$$

The mean square error of an estimate is a theoretical property based on the two unknown quantities, variance [V($\hat{\beta}$)] and bias [Bias ($\hat{\beta}$)]. A researcher using an estimate in a decision-making process will probably want the estimate to have the smallest mean square error rather than the smallest of one of its two components.

The choice of estimation procedure depends crucially on the prospective use of an estimate. In some cases the researcher looks for unbiased estimates at any cost, in other cases he wants maximum precision even though the estimate may be biased, in yet other cases he wants the minimum mean square error. The point is that in applied econometrics the researcher should be alert to all aspects of his estimation procedure and should not rely on only one of them.

2.5 Best Linear Unbiased Estimates

In a typical textbook example of a regression equation, the causal relationship between the dependent and independent variables is linear and the specified relation is the truth. In this case the independent variables are clearly defined and exhaust the sources of variation in the dependent variable. We may express the true specification as

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_k x_{kt} + \varepsilon_t \ , \qquad (2.37)$$

where lower-case letters represent the deviations of the variables from their respective means. In this case the constant term is implicit.

Suppose the researcher conducts a hypothetical experiment in which he draws a set of error terms ($\varepsilon's$) at random from a known distributions and computes *y* on the basis of these errors and of the independent variables (*x*'s). Corresponding to a set of errors he obtains a set of least squares estimates for the β's. If he repeats the experiment, drawing a new set of errors every time but keeping the same values for the *x*'s, then the least squares estimates of the β's obtained in each trial will exhibit a statistical distribution.

The Gauss-Markov theorem states that when the estimated equation is the truth, the distribution of the ordinary least squares estimate of any of the parameters in the above experiment is centered around the true values; that is, all the estimates are unbiased. Further, the ordinary least squares estimates are the least variance estimates in the family of linear unbiased estimates.

When an estimate is unbiased and has minimum variance in the family of linear unbiased estimates, then it is called the Best Linear Unbiased Estimate (BLUE), or the minimum variance unbiased estimate. This is a theoretical property of the estimation procedure because the bias and variance are based on the theoretical statistical distribution of the least squares estimate in the above experiment.

In applied econometric research, the theoretical properties of an estimate are unattainable goals and serve only as points of reference. Researchers seldom know the truth which would let them take advantage of any of these theoretical properties.

When an applied econometrician deviates from the truth—for example, by estimating a regression equation which is not the truth or by using a definition of a variable that is not the truth—he commits a "sin." That is, his estimates may not have the same theoretical properties as though he had used the truth. In all real situations the alternatives open to an applied econometrician involve certain amounts of such sin, and the best he can hope for in estimating a parameter is to follow the least sinful procedure of them all.

2.6 Left-Out Variables

The true regression specification can be estimated only when the researcher knows the truth and has data on all the variables to estimate it. A common situation is one in which the researcher has "left out" variables either because he is unaware of their presence in the true specification or because he does not have data for including them in the estimated equation.

The use of ordinary least squares when some variables are left out may introduce bias into the estimates. Bias that originates in this way is called specification bias. For instance, let the truth be

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t . \qquad (2.38)$$

Here the lower-case letters are again deviations from the mean, hence the constant is implicit. We assume that no data are forthcoming for the variable x_2, which is therefore operationally unobservable, so the following regression equation is estimated:

$$y_t = \hat{\beta}_1 x_{1t} + e_t . \qquad (2.39)$$

The ordinary least squares estimate $\hat{\beta}_1$ is obtained by imposing restrictions on the residuals (e's). Since the constant term is implicit in the estimation, $\Sigma e = 0$. The second constraint ($\Sigma x_1 e = 0$) yields

$$\Sigma x_1 y = \hat{\beta}_1 \Sigma x_1^2 . \qquad (2.40)$$

The estimate $\hat{\beta}_1$, that satisfies the constraints on the residuals is

$$\hat{\beta}_1 = \Sigma x_1 y / \Sigma x_1^2 . \qquad (2.41)$$

Since the truth is given by (2.38), equation (2.41) can be rewritten as

$$\hat{\beta}_1 = (\beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_1 x_2 + \Sigma x_1 \varepsilon) / \Sigma x_1^2 . \qquad (2.42)$$

Since the *x*'s are held constant in repeated trials and the distribution of ε is assumed to have zero mean, the mean value of the theoretical distribution of $\hat{\beta}_1$ is

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \Sigma x_1 x_2 / \Sigma x_1^2 . \qquad (2.43)$$

The estimate of β_1 from the ordinary least squares estimation of (2.40) when the truth is (2.38) is a biased estimate of the parameter β_1. The bias ($\beta_2 . \Sigma x_1 x_2 / \Sigma x_1^2$) depends on two terms, namely the regression coefficient of the left-out variable in the true relation (β_2), and the comovements of the left-out variable with the included variable ($\Sigma x_1 x_2 / \Sigma x_1^2$).

The expression in equation (2.43) can be generalized into a case with *K* variables by using the Yule notation. A linear regression equation in Yule notation may be written as

$$y = b_{y1.23\ldots k} x_1 + b_{y2.13\ldots k} x_2 + \ldots + b_{yk.123\ldots k-1} x_k + e . \qquad (2.44)$$

The coefficients (*b*'s) are given subscripts in a systematic way. The first subscript denotes the dependent variable, the second denotes the corresponding independent variable. The list after the comma (,) indicates other independent variables included in the regression. Since the estimates of the regression coefficients depend on which other independent variables are present in a regression, these subscripts are explicitly stated in each case.

In the case with x_2 as the dependent variable and x_1 as the independent variable, the regression equation in the Yule notation may be written as

$$x_2 = b_{21} x_1 + e . \qquad (2.45)$$

Where b_{21} is the ordinary least squares estimate from equation (2.45):

$$b_{21} = \Sigma x_1 x_2 / \Sigma x_1^2 . \qquad (2.46)$$

Using the Yule notation we may rewrite expression (2.43) as

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 . b_{21} , \qquad (2.47)$$

where b_{21} is computationally equivalent to the regression coefficient when x_2 is the dependent variable, x_1 is the independent variable, and no other variables are present in the regression equation. This term is used in simplifying the algebraic expressions and does not have any causal or economic interpretation.

The expression for bias in the general case where the truth is

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_k x_{k1t} + \varepsilon_t \ , \qquad (2.48)$$

and where the following regression equation, without x_2, is estimated,

$$y_t = \hat{\beta}_1 x_{1t} + \hat{\beta}_3 x_{3t} + \ldots + \hat{\beta}_k x_{kt} + e_t \ , \qquad (2.49)$$

the expected value of the estimate $\hat{\beta}_1$, is given by

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 . b_{21.3\ldots k} \ , \qquad (2.50)$$

where $b_{21.3\ldots k}$ is computationally equivalent to the regression coefficient of x_1 when x_2 is the dependent variable and all the variables $x_1, x_3, \ldots, x_k$ are included in the regression equation.

When an independent variable in the true relation is omitted, the regression coefficients from the OLS estimation procedure are biased. The extent of bias in each coefficient can be obtained from equation (2.50). When a variable from the true relation is left out, a part of its influence in explaining the movements of the dependent variable is captured by the other independent variables. The relative share of each included variable in capturing the influence of the left-out variable is given by the *b*'s (also called the auxiliary regression coefficients). If one independent variable has a larger partial relation to the left-out variable than another, then the extent of bias in its coefficient will be larger.

When the left-out variable is not correlated with any of the independent variables then, of course, none of the coefficients is biased. In any sample the researcher rarely observes zero correlation; hence some bias always exists, however small it may be. The applied econometrician is not worried about the mere existence of bias, but about its extent. When the bias is of second order in magnitude (smaller than the rounding error in truncating decimals), or even smaller, it causes no concern in most practical situations.

2.7 An Example of Bias

Consider the linear regression equation (2.51) which attempts to explain the quantity of rice produced in the Guntur district of India for the period 1941-61.

$$\begin{array}{l} \text{Rice} = 993.633 + 0.046\ \text{I} + 0.706\ \text{D} + 48.219\ \text{R} \qquad R^2 = 0.56 \qquad (2.51) \\ \qquad\quad (1368.440)\ \ (0.273) \quad (0.945) \quad (11.282) \end{array}$$

where *I*, *D*, and *R* are acres of irrigated area, acres of dry area, and inches of rainfall respectively.

The estimates of the regression parameters are disturbing, because it is well known that dry land does not produce 0.706 tons at the margin whereas irrigated land produces only 0.046 tons at the margin, keeping all other independent variables constant. This strange result is a consequence of misspecification of the estimated regression equation. The true specification of the rice production function includes many variables in addition to those included in (2.51).

When the nature and data of these other variables are known, then we will include them in the regression equation to correct for the bias in (2.51)

In this particular example, let us assume that data on the other variables are not available. We may conjecture that the combined influence of the variables left out in the true equation is a smooth function of time. This influence could comprise any systematic factors that affect the production function, whose movements may be changing with time in a smooth linear form.

The regression equation, with "*t*," time, as an explicit independent variable for our example, is

$$\begin{array}{l} \text{Rice} = -739.9\ + 0.578\ \text{I}\ + 0.218\ \text{D} + \ 46.6\ \text{R}\ - 40.4\ \text{t} \qquad R^2 = 0.61 \qquad (2.52) \\ \qquad\quad (1755.4)\ \ (0.442) \quad (0.959) \qquad (10.9) \quad (26.9) \end{array}$$

Equation (2.52) is consistent with our a priori experience. Comparison of (2.51) with (2.52) reveals that the omission of a crucial variable from equation (2.51) has caused the independent variables, *I* and *D*, to capture a part of the omitted variable's influence on the dependent variable. In the case of *I* the bias is negative, and in the case of *D* it is positive. Apparently the left-out variable has not substantially biased the coefficient of the rainfall variable.

In this particular example, equation (2.51) indicates the presence of specification bias, and equation (2.52) sheds some light on the nature of the problem. The variable "*t*" here helps us by pointing out the presence of specification bias but does not explain what variables caused it. When the researcher has knowledge of the variables left out, inclusion of them will improve the situation. In our example, one has to explore for the factors causing the specification bias by using the signals provided by the variable "*t*". In many practical problems the bias in the estimates may not be so conspicuous.

2.8 Irrelevant Variables

A case inverse to that of left-out variables is the case in which a variable not specified in the true equation is added. Such variables are called irrelevant.

Consider the true equation

$$y_t = \beta_1 x_{1t} + \varepsilon_t . \qquad (2.53)$$

Instead, the following equation is estimated:

$$y_t = \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} + e_t . \qquad (2.54)$$

The OLS estimate $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \frac{\Sigma x_2^2 . \Sigma x_1 y - \Sigma x_2 x_1 . \Sigma x_2 y}{\Sigma x_1^2 . \Sigma x_2^2 - \Sigma x_1 x_2 . \Sigma x_1 x_2} . \qquad (2.55)$$

The true relation is given by the equation (2.53). By substituting this expression in (2.55) we obtain

$$\hat{\beta}_1 = \beta_1 + \frac{\Sigma x_2^2 . \Sigma x_1 \varepsilon - \Sigma x_1 x_2 . \Sigma x_2 \varepsilon}{\Sigma x_1^2 . \Sigma x_2^2 - \Sigma x_1 x_2 . \Sigma x_1 x_2} . \qquad (2.56)$$

When expected values are taken on both the sides of (2.56), it may be seen that the OLS estimate of β_1 from equation (2.54) is nonetheless an unbiased estimate: $E(\hat{\beta}_1)=\beta_1$.

The addition of an irrelevant variable to a true specification does not cause bias in the estimates of the other independent variables. However, such an addition does necessarily increase the variance of the estimates of all the coefficients. Thus, even though the

irrelevant variable does not introduce bias in the regression coefficients, it will reduce the precision of the estimates. This result can be generalized to a *K*-variable model.

The researcher may verify that the ordinary least squares estimate of β_2 is also an unbiased estimate. The mean value of the theoretical distribution of $\hat{\beta}_2$ is centered around its true value, namely zero. For analytical convenience the true equation (2.53) may also be written as

$$y_t = \beta_1 x_{1t} + 0.x_{2t} + \varepsilon . \qquad (2.57)$$

Obviously, in an empirical situation the addition of an irrelevant variable will typically yield a nonzero coefficient value for $\hat{\beta}_2$. This should not be interpreted as indicating that its distribution has a nonzero mean, for such is not the case when the true equation is (2.53).

2.9 Superfluous Variables

Up to this point we have been dealing with cases in which the truth is known. In applied econometrics the expressions for bias and variance offer only cold comfort, for in many practical situations it is neither simple nor unambiguous to state which is a left-out variable or an irrelevant variable. The theoretical results previously cited are useful only when we can “stick our necks out” and categorically state what we believe to be the truth.

When the choice between competing alternatives will not make serious operational difference, a practicing econometrician should follow guidelines that involve a certain amount of judgment; he should not rely solely on knowledge of theoretical derivations.

When a new independent variable is added to a regression equation, the regression coefficients either (1) change substantially or (2) do not change substantially. When the coefficients change, it can be either because specification bias was present in the regression coefficients before the variable was added or because the added variable is irrelevant and the estimates after inclusion of the new variable are coming from a different distribution with larger variance but with the same mean.

When the true equation is known we can distinguish between these possible situations, but when the researcher has only the computed regressions the problem requires a less straightforward assessment. Whenever the regression coefficients in the two situations are obviously different, our decision as to which equation to use may be extremely

crucial for the interpretation of the parameter estimates. In this case we must rely purely on the theoretical underpinnings of the regression equation to tell us which to choose.

On the other hand, we may have the following situation, in which the researcher fits the two regressions

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_{1t}, \qquad (2.58)$$

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_{2t}. \qquad (2.59)$$

Assume that no unambiguous theory exists that specifies which of the two equations is the truth. If equation (2.58) is true, estimation of equation (2.59) yields an unbiased estimate of β_1 but with a larger variance than to the corresponding estimate from equation (2.58). If equation (2.59) is true and (2.58) is estimated, then $\hat{\beta}_1$ is biased but has a smaller variance than the corresponding estimate from (2.59). Which equation should be chosen for empirically estimating the parameters is not easily answered, and the distribution properties of the estimates help us very little.

In some situations, the variable X_2 may be such that, if (2.58) is true, adding it as an independent variable and estimating equation (2.59) will yield less precise estimates, but the loss in precision may be minor. It may also happen the other way—that is, if equation (2.59) is the true specification, leaving out the variable X_2 and estimating (2.58), will yield biased estimates, but the bias may be minor. If X_2 displays these two traits when included and excluded respectively, then we label it as a superfluous variable. A superfluous variable's exclusion or inclusion in a regression equation does not alter the other coefficient estimates or their standard errors "significantly." That is, when such a variable is added or deleted in a regression, the "damage" to the regression coefficients of the other variables is not operationally significant. Since we do not know the truth, we use or discard the superfluous variables to suit our convenience.

Some standard guidelines in the use of superfluous variables are given here, but a word of warning is needed. These rules are applicable only to variables whose legitimacy in the regression under study is dubious. When the theory unambiguously states that a variable is a specified explanatory variable then, of course, it should not be omitted even though it might appear superfluous.

As one guideline, then: when X_2 is a superfluous variable and its inclusion does not increase R^2 sufficiently to increase the $\bar{R}^2$, then the variable is deleted from the regression. When inclusion of a superfluous variable increases R^2 sufficiently to increase

$\bar{R}^2$, then it is included because it reduces the residual variance but does not affect the other regression coefficients. When its inclusion reduces the residual variance, this implies that the residuals have some systematic component which is being captured by the included variable, and error terms are thereby better specified.

What we mean by "better specification of the errors" may be seen in the following example. Suppose the two possible truths are (2.58) and (2.59). When both are estimated, it is seen that the estimate of β_1 does not change when (2.59) is estimated but that $\bar{R}^2$ goes up. Here X_2 is superfluous, but ε_1 contains $\beta_2 X_2$ which is systematic, so that ε_2 is better specified in (2.59) as a random variable because the systematic part, $\beta_2 X_2$, was purged from ε_1. So we say that in this particular case we may choose (2.59) as the appropriate regression equation because (1) we can still correctly interpret the estimated results and (2) we have significantly taken account of a systematic part of the error term.

A common practice among many researchers is to compute the *t*-ratios of the regression coefficients defined as

$$t(\hat{\beta}_i) = \hat{\beta}_i / st.error(\hat{\beta}_i) \qquad (2.60)$$

and to discard the variables with *t*-ratios below a certain small value. This is not recommended as a common research method, for it is not possible to recognize a superfluous variable solely on the basis of the magnitude of its *t*-ratio. It is true that when a variable with a *t*-ratio smaller than unity is discarded from a regression equation the $\bar{R}^2$ increases, but this increase does not imply that the variable is superfluous. It may or may not be. If the discarding of a variable changes the regression coefficients of other variables, then it cannot be superfluous even if its deletion increases $\bar{R}^2$. In such cases even if the variables in question have very small *t*-ratios their continued presence in the regression equation is dictated by theoretical considerations and not by the summary statistics.

2.10 Example of Misuse of Criteria for Identification of Superfluous Variables

If the researcher starts out with conceptual errors, the standard criteria for identification of superfluous variables can no longer be used. For example, suppose theory has indicated that the demand for Ceylonese tea (TEA) in the United States is a function of disposable income (Y^d), the price of Ceylonese tea (P_{cy}), and the price of a close substitute—Brazilian coffee (P_{BZ}) all relative to the price of food commodities in the United States. In the log-linear form the estimated demand function is

$$\log \text{TEA} = \underset{(1.99)}{3.95} + \underset{(0.14)}{0.14} \log P_{BZ} + \underset{(0.24)}{0.75} \log Y^d + \underset{(0.41)}{0.05} \log P_{CY} \quad \bar{R}^2 = 0.52 \qquad (2.61)$$

Here the t-ratio of the price variable (P_{cy}) is very "low" and the coefficient has the "wrong" sign. If the researcher concludes that the demand for Ceylonese tea is price inelastic and discards (P_{cy}) while reestimating the demand function, he gets

$$\log \text{TEA} = \underset{(0.71)}{3.73} + \underset{(0.13)}{0.14} \log P_{BZ} + \underset{(0.14)}{0.73} \log Y^d \quad \bar{R}^2 = 0.54 \qquad (2.62)$$

By comparing (2.61) and (2.62) he may jump to the conclusion that the price variable (P_{cy}) is superfluous and might interpret it as evidence that the demand for Ceylonese tea is, in fact, price inelastic.

He will be misguided, though, because he has forgotten one salient aspect of the demand for Ceylonese tea in the United States: the strong competition from other tea-growing countries of the world. He should be reluctant to accept the implication that the price of other close substitutes does not influence the demand for Ceylonese tea.

One must, therefore, always search the theory for any conceptual errors. In this case the error can be readily recognized, since Ceylonese tea and Indian tea are such close substitutes in consumption that as the price of Indian tea changes, so does the demand schedule for Ceylonese tea. The latter can be properly interpreted only when the price of Indian tea is held constant.

Multiple regression analysis allows us to hold other specified variables constant only when they are included in the regression equation. Clearly, the demand for tea equation (2.61) did not include the price of Indian tea, and therefore the coefficient of log P_{cy} was including two effects: changes in the price of Ceylonese tea (movements along a given demand schedule) and changes in the price of Indian tea (shifts in the demand schedule).

Even if our coefficient estimate for log P_{cy}, were of the correct sign, we would still be observing the combined effect, and as a result the variable might seem superfluous according to our usual criteria.

Inclusion of the price of Indian tea (P_I) in the regression equation gives the following estimates:

$$\log \text{TEA} = \underset{(2.00)}{2.84} + \underset{(0.13)}{0.19} \log P_{BZ} + \underset{(0.37)}{0.26} \log Y^d - \underset{(0.98)}{1.48} \log P_{CY} + \underset{(0.69)}{1.18} \log P_I \quad \bar{R}^2 = 0.56$$

The empirical results now confirm our theoretical expectations. A general rule, then, is not to throw out automatically a variable which seems superfluous by standard criteria. Rather, one must first fully eliminate all possible conceptual errors that could have led to these results. When there is a conceptual error one must not look to his data to tell him whether a superfluous variable is present in his regression equation.

2.11 Dominant Variables

In many empirical studies theory tells us unambiguously that a particular independent variable is relevant in explaining the movements of the dependent variable, but it cannot be included in the regression equation because it is a dominant variable. Such a variable dominates all the other variables in the regression and attempts to account for all variation in the dependent variable, leaving nothing to be explained by other variables. This situation frequently occurs in empirical research where the dependent variable is somehow functionally related to an independent variable in "fixed proportions." For example, in a production function of wheat a relation is fixed between the output and the amount of seed used. This relation is fairly constant for all observations, and when the amount of seed is used as an independent variable it explains all the variation in the dependent variable despite the presence of other variables such as capital, labor, etc.

Dominant variables present a problem of estimation techniques rather than a theoretical issue as to the relevance of an independent variable. As E. D. Domar puts it:

> I wonder what has happened in all these [production function] studies to material inputs. If they are omitted because of the lack of required data, we have an answer, even if, to my mind, a regrettable one .... It seems to me that a production function is supposed to explain a productive process, such as the making of potato chips from potatoes (and other ingredients), labor and capital. It must take some ingenuity to make potato chips without potatoes.

Domar's statement cannot be contested on theoretical grounds. Empirical work, however, may require the deletion of primary material inputs in a production function simply because these inputs are so "dominant" as to cause all other included variables to become superfluous. An example will serve to demonstrate this point.

Consider the problem of estimating the production function for Indian woolen textile industries. Let the production function, relating the output (Q) with the inputs—capital

(*K*), labor(*L*), and raw materials (*M*)—be

$$Q = A K^{\beta_1} L^{\beta_2} M^{\beta_3} e^{\varepsilon} \quad , \qquad (2.64)$$

where the error terms are specified as exponential to the natural base (e).

When the prices of all the inputs and of output are constant for all observations the production function may also be expressed as

$$q = \alpha k^{\beta_1} l^{\beta_2} m^{\beta_3} e^{\varepsilon} \quad , \qquad (2.65)$$

where the lower-case letters now stand for the values of output and of inputs. The parameters (β's) in equations (2.64) and (2.65) are the same.

The production function for woolen textiles estimated from the data of Indian Woolen Textiles is

$$\log q = \underset{(0.256)}{-0.408} \; \underset{(0.043)}{-0.059 \log k} \; \underset{(0.051)}{-0.002 \log l} \; \underset{(0.065)}{+1.094 \log m} \qquad \bar{R}^2 = 0.997 \qquad (2.66)$$

In a production function we expect all the β's to be positive, but the regression equation (2.66) yields negative values for the estimates of β_1, and β_2. By discarding the variables (log *k*) and (log *l*) and re-estimating the production function we obtain

$$\log q = \underset{(0.232)}{-0.531} \; \underset{(0.014)}{+1.049 \log m} \qquad \bar{R}^2 = 0.998 \qquad (2.67)$$

A comparison of equations (2.66) and (2.67) leads to the conclusion that capital and labor are superfluous variables in the production of woolen Textiles. This, however, is a misinterpretation of the regression equation.

In this example the raw materials variable is dominant, and is being transformed into output at a more or less fixed conversion ratio. The inclusion of log *m* as an independent variable leaves no room for other inputs. By deleting this variable (log *m*) from regression equation (2.66) and re-estimating the function, we obtain

$$\log q = \underset{(1.255)}{-0.026} \; \underset{(0.161)}{+0.413 \log k} \; \underset{(0.138)}{+0.708 \log l} \qquad \bar{R}^2 = 0.938 \qquad (2.68)$$

Estimates of β_1, and β_2 in equation (2.68) have proper signs and reasonable magnitudes.

Even though, theoretically speaking, the production function is given by equation (2.65), its parameters are not empirically estimable, because of the dominant role played by the input (raw materials). The production functions given by (2.66) and (2.68) may not have the same theoretical interpretation. However, the regression equation (2.68) can legitimately be interpreted as describing a production process that uses capital and labor in transforming raw materials into output. The output may be viewed either as the product (textiles) or the amount of raw material (wool) processed by the industry. As long as there is one-to-one correspondence between the raw materials and the product, the production function given by the regression equation (2.68) is operationally useful—for example, in hiring and firing labor—and has not been harmed by the exclusion of raw materials as an input.

Some researchers have overcome the conceptual problems presented by raw materials by deleting them and specifying the dependent variable in a production function as value added by manufacturing instead of total real output or as the total value of output.

The presence of dominant variables cannot be ascertained on a priori grounds. There may always be a possibility of substitution between the dominant variable and the other independent variables. For example, consider the production of wheat. It might appear that a farmer must always plant the same proportion of seed to wheat. Yet some farmers may be able to obtain the same yield by planting less seed while employing extra workers to scare the crows away so that none of the seed will be eaten before germination. Other farmers may find it more economical to let the crows have their way, not to hire extra workers, and instead to plant more seeds. Here labor can be substituted for raw materials. We cannot know that the substitution took place until we attempt to estimate our wheat production function. Only in the case in which some farmers substitute labor for seed and others do not, can seed be treated as a nondominant variable in the regression equation. That is, if the sample shows a sufficient amount of substitution between the raw material and other variables, then the raw material will not show up as dominant.

A similar example occurs in the production of sugar, in which substitution is possible between capital used and sugar cane. With more capital, one can extract a larger percentage of sucrose from the cane.

Whether a variable is truly superfluous or is a consequence of the presence of dominant variables must be determined by investigation, and no rule of thumb can be given to solve this problem. If conceptual errors exist in the specification, these guidelines are

not valid. It is to be kept in mind that no statistical tool or econometric guide is a good substitute for theory. Guidelines indicate where to look in case of trouble, but not necessarily how to solve the problem.

2.12 Regression Coefficients with Wrong Signs

Often when regression coefficients are estimated the sign is opposite to that which the researcher believes to be true. When this happens, many researchers unfortunately drop the guilty variable from the regression equation with no further mention. In many cases, however, this is not an acceptable procedure, for a wrong sign may be a warning, inter alia, of incorrect definitions, specifications, or interpretations.

We frequently interpret coefficients in such a way that the estimated sign or magnitude appears to be different from what is expected. This may be due to an incorrect interpretation, which in turn may have resulted from treating an implicit form of the equation as the explicit form. That is, our estimating equation may be derived from another equation; and thus some of our explicit coefficients may be equivalent to a function of some unrecognized implicit coefficients.

Consider, for example, the following income determination equation for India for the period 1951-63:

$$\underset{\quad (11.42)\quad (0.455)\quad (0.147)}{Y_t = 11.20 + 0.406\, I_t + 0.887\, Y_{t-1}} \qquad \bar{R}^2 = 0.96 \qquad (2.69)$$

where Y is the national income in constant prices and I is investment in constant prices defined as the sum of savings and net capital inflow (Reserve Bank of India estimates).

If the researcher interprets the coefficient of I as the investment multiplier, he will immediately conclude that there is a misspecification of some sort. If he recomputes the regression by discarding the variable Y_{t-1}, he obtains the following equation:

$$\underset{\quad (4.40)\quad (0.377)}{Y_t = 77.27 + 2.997\, I_t} \qquad \bar{R}^2 = 0.84\,. \qquad (2.70)$$

The coefficient of I in (2.70) is consistent with the a priori notions of the investment multiplier.

If comparison of (2.69) and (2.70) then leads to the conclusion that Y_{t-1} does not belong in the equation, the deduction is wrong. In the first place, the coefficient of I_t in (2.69)

should not be interpreted as the long-run investment multiplier; rather, it is the impact investment multiplier. The coefficient of I_t in (2.70) also cannot be interpreted as the long-run investment multiplier, for equation (2.70) violates economic theory: it implies that investment has a once-and-for-all one-shot effect on income.

If the researcher desires the true long-run multiplier, he must look at the steady-state solution of (2.69), where Y_t, is assumed to be equal to Y_{t-1}. He then obtains an implicit long-run investment multiplier of

$$0.406 / (1 - 0.887) = 3.6.$$

Had the researcher understood the implications of (2.69), his remarking of a "small" coefficient of I_t would not have induced him to throw Y_{t-1} out of the equation.

Often the estimated results need not be conspicuous to warn of the presence of conceptual errors. When a signal is given in the form of a wrong sign or magnitude, the researcher should not ignore the warning, but rather should try to improve his specification of the regression equation.

A common cause of wrong signs in empirical research occurs when the variables are not appropriately defined. Consider, for example, the demand for a commodity (Q) as a function of disposable income (Y), the price of the commodity (P), and the price of all other commodities represented by the wholesale price index (W):

$$\log Q_t = \beta_0 + \beta_1 \log Y_t + \beta_2 \log P_t + \beta_3 \log W_t + \varepsilon_t \ . \qquad (2.71)$$

Suppose the variable Y is measured in current prices, whereas the true relation is in constant prices. If disposable income in constant prices is (Y/W), then the true relation is

$$\log Q_t = \alpha_0 + \alpha_1 \log (Y/W)_t + \alpha_2 \log P_t + \alpha_3 \log W_t + \varepsilon_t \ , \qquad (2.72)$$

which simplifies to

$$\log Q_t = \alpha_0 + \alpha_1 \log Y_t + \alpha_2 \log P_t + (\alpha_3 - \alpha_1) \log W_t + \varepsilon_t \ . \qquad (2.73)$$

The estimated coefficient of log W in equation (2.73) is actually a function of several parameters, and interpretation as if it were $\hat{\alpha}_3$ leads to wrong conclusions.

If specification and interpretation of the coefficients are correct, a coefficient can still attain a wrong sign because of the sampling distribution of the estimates. If this is the case, we generally observe the coefficient to be not significantly different from zero

statistically; deletion of the variable because of its wrong sign may still lead to misspecification. When the coefficient is significantly different from zero statistically and has the wrong sign, then some aspect of the problem has not been unveiled. Instead of throwing away the variable, it is better to retain it in the equation so that other researchers may be able to explain the apparent inconsistency. When the variable with a wrong sign is superfluous, in the sense that its deletion does not affect the other coefficients and does not decrease $\bar{R}^2$, then this problem is not serious.

2.13 Multicollinearity

In some cases, even though a theory clearly indicates the independent variables in explaining movements in a dependent variable, it may not be possible to interpret some of the parameters of the regression equation.

Consider the following example:

$$S_t = \beta_0 + \beta_1 L_t + \beta_2 R_t + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t \ , \qquad (2.74)$$

where *S* is sales revenue, *L* is the number of left shoes sold, *R* is the number of right shoes sold, and *X3* and *X4* are other products. The revenue comes from selling both the right and the left shoes, therefore each has legal claim in explaining the movements in sales revenue; but then some of the parameters in equation (2.74) have no meaningful interpretation.

The parameter β_1, for example, is the partial derivative of *S* with respect to left shoes, keeping all the other variables, including right shoes, constant. Such a situation is never observed because shoes are always sold in pairs. Even if the parameter values β_1, and β_2 in (2.74) were somehow obtained, they could not be interpreted. This problem arises whenever there is a fixed relationship between independent variables.

In many empirical problems the basis of interconnection may not be so conspicuous, but even in the absence of any theoretical reason the data may be such that a one-to-one relation between the variables is present. Whatever the source of the fixed relation, the problem can be averted by redefining the variables in such a way as to make the parameters subject to interpretation.

For example, in the above case instead of specifying the independent variables as left and right shoes separately, a new variable defined as “a pair of shoes” may be used. The equation then becomes

$$S_t = \beta_0 + \beta P_t + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t \,, \qquad (2.75)$$

where *P* is the number of pairs of shoes sold.

Now consider the reverse case. In many situations, even though theoretically a technical relation exists between independent variables, the observed data may not exhibit any such relation. Consider, for example, the following hypothetical equation to explain long-run price levels:

$$P_t = \beta_0 + \beta_1 C_t + \beta_2 DD_t + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t \,, \qquad (2.76)$$

where *P*, *C*, and *DD* are price levels, currency, and demand deposits, respectively. Suppose a law requires a minimum currency reserve ratio (*C/DD*). The researcher may expect a technical relation between *C* and *DD*; but if the banks in fact maintain free reserves he does not observe any such technical relation in his data. In such cases the parameters of equation (2.76) do have a valid interpretation. Regardless of what might have been expected a priori, the technical relation between independent variables poses problems of interpretation only when it appears in the data.

The presence of any fixed relation between independent variables presents a problem called multicollinearity.

Although researchers show a growing tendency to blame all econometric problems on this demon, we suggest that it may often be largely a theoretical nightmare rather than an empirical reality.

The applied econometrician does need some guidelines in order to detect the presence of multicollinearity in his data, however, and a few indicators are available. A standard rule that some investigators have been using calls for inspection of the simple correlations among the independent variables. One should realize that simple correlations are only elements of the entire correlation matrix and, hence, may or may not contribute to problems of multicollinearity. One should not, a priori, rule out estimation of any regression equation because of high simple correlations between any two independent variables.

Consider, for example, the following regression, which attempts to explain heartbeat (*Y*) among cardiac patients in a given hospital by the length of the right leg (*X1*) and the length of the left leg (*X2*):

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t \ . \qquad (2.77)$$

If the researcher computes the simple correlation between X1 and *X2* he will observe "high" correlation; he therefore suspects multicollinearity and hesitates to estimate the equation. If another researcher were to look at the same equation with *X1* and X2*(= *X2* - *X1*) he would see a "low" correlation between *X1* and X2* and conclude that multicollinearity was not a problem. But the implicit estimates of β_1 and β_2 from (2.77) will be identical whether the second independent variable is *X2* or *X2**. This is so because the conditions imposed on the residuals for estimation in either case are implicitly the same.

The researcher would get a meaningful regression equation by including *X1* and *X2*. If one or the variables were deleted, he would obtain nonsensical results. Because of the asymmetry of the heart, its beat is a function of difference in the length of the legs and not of the length of either. By including both variables, the researcher is implicitly allowing the difference in the two variables to enter the regression equation as an independent variable.

When one independent variable is a linear function of the other, then the ordinary least squares estimation procedure fails. For example, in equation (2.77) the estimates of the three parameters β_1, β_2 and β_3 are obtained by solving the three constraints on the residuals:

$$\Sigma e = 0 \ , \qquad (2.78)$$
$$\Sigma x_1 e = 0 \ , \qquad (2.79)$$
$$\Sigma x_2 e = 0 \ . \qquad (2.80)$$

A nontrivial solution exists only when the above three constraints are independent. When a linear relation exists between *X1* and *X2*, the second and the third equations are not independent. Hence, when *X2* is a linear function of *X1* the ordinary least squares solution of $\hat{\beta}$'s cannot be obtained. This property may be used in detecting the existence of fixed relations between independent variables.

When a regression equation has several parameters to be estimated, the least squares algorithm for $\hat{\beta}$ is usually expressed in matrix notation as

$$\hat{\beta} = (X'X)^{-1}X'Y \ . \qquad (2.81)$$

When a linear relation exists between independent variables, then the matrix (*X'X*) is singular and has no inverse. The applied econometrician rarely inverts the matrix in one step. All algorithms obtain the inverse by "pivoting- in" (including) one variable at a

time. When a variable being pivoted-in is a linear function of already pivoted-in variables, the diagonal element of the matrix becomes zero, or nearly zero, when computational errors are present. Such variables can be detected and skipped with no difficulty. Most computer programs check for zero elements along the diagonal at each step and notify the researcher of the presence of any linear relation. These problems can easily be remedied by appropriately redefining the independent variables.

Sometimes the computational errors and errors of measurement can be large enough to cause a nonzero diagonal element in the matrix. In such cases the regression coefficients will remain fairly stable before and after the variable is introduced, but the standard errors of all the coefficients will increase beyond reasonable limits following its introduction. When X_1, and X_2 are collinear, X_2 will turn out to be superfluous when X_1 is already in the regression, and vice versa. The $\bar{R}^2$ invariably goes down when the second variable is added to the equation.

Some researchers attribute even a slight change in standard errors to multicollinearity. This practice should be discouraged. Consider, for example, the two estimated equations

$$y_t = \hat{\beta}_1 x_{1t} + e_{1t} , \qquad (2.82)$$
$$y_t = \tilde{\beta}_1 x_{1t} + \tilde{\beta}_2 x_{2t} + e_{2t}, \qquad (2.83)$$

where both were estimated by ordinary least squares.

The estimates $\hat{\beta}_1$ and $\tilde{\beta}_1$ have two different theoretical distributions with different variances (see p. 56):

$$V(\hat{\beta}_1) = \sigma_\varepsilon^2 / \Sigma x_1^2 , \qquad (2.84)$$
$$V(\tilde{\beta}_1) = \sigma_\varepsilon^2 / \Sigma x_1^2 (1 - r_{x1x2}^2) . \qquad (2.85)$$

When correlation between the two independent variables (r_{x1x2}) differs from zero, then the variance of $\tilde{\beta}_1$ is larger than that of $\hat{\beta}_1$.

In some situations the correlation between the independent variables may be close enough to unity to make the variance of the estimate extremely large. In such cases the estimate is not “reliable.” When the (unknown) standard deviation of the estimate is very large relative to the mean of the distribution, the researcher may discard the variable to improve the mean square error of the other estimates. He should remember that the theoretical variance of an estimate depends not only on the correlation between the independent variables but also on the variation of the independent variables; for example, the variances of both the estimates $\hat{\beta}_1$ and $\tilde{\beta}_1$ decrease with Σx_1^2.

Even though the correlation between the independent variables is "nearly" unity, the variation in the independent variable Σx_1^2 may offset the term (1- r_{x1x2}^2) and make the theoretical variance very small. Since the value of Σx_1^2 increases with the sample size, the problem of multicollinearity, except in the sense of a fixed relation between the independent variables as in the example of right and left shoes, does not usually arise in large samples.

When multicollinearity is present, the exclusion of one of the variables from the regression equation does not decrease the explanation of the dependent variable.

Consider the equation

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t .$$

Let there be a linear functional relationship between X_1 and X_3

$$X_{3t} = a + b X_{1t} . \qquad (2.87)$$

Suppose that instead of (2.86) the following regression equation is estimated:

$$Y_t = \hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \hat{\beta}_2 X_{2t} + e_t . \qquad (2.88)$$

By using equation (2.50), we have

$$E(\hat{\beta}_1) = \beta_1 + \beta_3 . b , \qquad (2.89)$$
$$E(\hat{\beta}_2) = \beta_2 . \qquad (2.90)$$

Since X_3 can be defined in any arbitrary units, we define the units such that $b = 1$, and we have equation (2.89) as

$$E(\hat{\beta}_1) = \beta_1 + \beta_3 . \qquad (2.91)$$

The entire influence of the variable X_3 is captured by the included variable, and the other coefficients are totally unaffected.

When the researcher is interested in explaining the movements of the dependent variable, or in predicting the values of *Y*, then it makes no difference whether the variable X_3 is in the regression or not. When his objective is to estimate the coefficients of the other independent variables, for example β_2, then again exclusion of the variable

X_3 will not damage the estimate. To isolate the influences of X_1 and X_3 however, becomes a computationally impossible task. This is a fortuitous situation, for even if the researcher were given the empirical estimates he would have no way of interpreting them.

The problem of which variables are to be included in a regression equation is a major problem in applied econometrics. Rules are helpful, but they cannot make decisions for the applied econometrician.